Chapter 21

Enhancing the Sentiment Classification Accuracy of Twitter Data using Machine Learning Algorithms

Muthukumar Bhuvaneswari[1] and Vasudevan Srividhya[2]

Abstract
Sentiment analysis or opinion mining is the study of public opinions, sentiments, attitudes, and emotions expressed in social media. This is one of the most dynamic research areas in natural language processing and text mining in current years. It is a domain that involves the finding of user sentiment, emotion and opinion within natural language text. The growing significance of sentiment analysis coincides with the increase of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. Common applications of sentiment analysis include the automatic determination of whether a review posted online (of a movie, a book, or a consumer product) is positive or negative toward the item being reviewed. This research work shows the various pathways to perform a computational treatment of sentiments and opinions. The main aim of this work is to classify the sentiment of twitter data using machine learning algorithms. The sentiment classifications have been classified into two types which are emotional classification and polarity classification. This work has been carried out on polarity classification, which is used to classify the text such as positive, negative, and neutral. The polarity classification is done by using the subjectivity lexicon. After the polarity classification two machine learning algorithms are employed to enhance the accuracy of sentiment classification. In the Pre-processing phase, the tweets are preprocessed by using various techniques. Sentiment classification is the essential phase, where preprocessed tweets are taken as input to sentiment classification. The sentiment classification can be done by using subjectivity lexicon. The third phase of the proposed work is to compare and evaluate the performance of two machine learning algorithms which are Support Vector Machine and Decision tree.

Keywords: Sentiment Classification, polarity, twitter, Support Vector Machine, Decision Tree.

---

[1]*M.Bhuvaneswari M.Phil, Department of Computer Science, Avinashilingam Institute for home science and Higher Education for Women, Coimbatore, Tamilnadu, India.*
[2]*Dr. V.Srividhya, Assistant Professor(SS), Department of Computer Science, Avinashilingam Institute for home science and Higher Education for Women, Coimbatore, Tamilnadu, India.*

S. Vignesh and A. Philip Arokiadoss (ed.), *Statistical Approaches on Multidisciplinary Research,* Volume I

Introduction

A large amount of data is generated in daily life due to the increased number of Social media network users. Data is available in different formats like text, audio, video, image and graphs. But the most important and the basic format that is used from the past till today is “Text”. Text plays a major role in communication.

Text Mining

Text mining is useful for handling textual data (NingZhong, 2012). Most of the textual data are unstructured, difficult to manipulate and unclear, so that text mining becomes the most useful method for information exchange whereas data mining is basically applied on business data (www.cis.upenn.edu). Text mining belongs to a non-traditional information retrieval strategy. The main goal of this approach is to reduce efforts required for obtaining information from huge set of textual documents.

Text Mining Tasks

Text Mining is part of data mining technique, can be used for various text related tasks such as concept/entity extraction, sentiment analysis, document summarization, information extraction, entity relation modeling (i.e., learning relations between named entities), categorization/classification and clustering.

Sentiment analysis or Opinion mining refers to a broad challenging area of NLP, computational linguistics and text mining. Aim of sentiment analysis is to determining the attitude of a speaker or a writer with respect to some topic. The attitude may be their judgment or evaluation, their affective state are the emotional state of the author when writing or the intended emotional communication which is the emotional effect the author wishes to have on the reader. Sentiment analysis task of text mining has gained importance because of the rise of social media such as blogs and social networks (www.wikipedia.org).

Sentiment classification is used to classify the opinion of a particular topic or product. Using this classification techniques identifying and predicting polarity becomes highly complex. The following are polarity classification,

- **Positive:** These are the good words about the target in consideration. If the positive sentiments are high, it is referred to be good.
- **Negative:** These are the bad words about the target in consideration. If the negative sentiments are high, it is discarded from the preference list. In case of reviews, if the negative reviews about the emotion are more, no one would intend to this option.
- **Neutral:** These are neither good nor bad words about the target. Hence it is neither preferred nor neglected.

Positive opinion words are used to express some optimistic states while negative opinion words are used to show some undesired state. Opinion lexicon is a collection of opinion phrases and idioms (Deebha Mumtaz, 2016).

The sentiment Analysis is widely applied in the social media networks. The most popular social networking websites are Face book, LinkedIn, Twitter and MySpace where people can communicate with each other by joining different communities and discussion groups. Twitter has become a huge social media service where millions of users contribute on a daily basis. Twitter is one of the evolving social media which, on an average, hosts around 200 million tweets per day. The proposed research work is carried out using twitter data.

Materials and Methods

The sentiment classification is used to identify the opinion of a particular topic or product. To classify the sentiment various classifications algorithms are used employed. “Sentiment Analysis and Measuring Opinions”, Chetashri Bhadane et al. (2015) implemented a set of techniques for aspect level sentiment classification of product review. Combination of both the machine learning algorithm which is Support Vector Machine(SVM) and domain specific lexicon are used for sentiment classification. The Support Vector Machine achieved 73% accuracy in the research paper. S.Veeramani et al.(2014) Web and social network, large amount of data are generated on Internet every day. This web data can be mined and useful knowledge information can be fetched through opinion mining process. This paper discussed different opinion classification and summarization approaches, and their outcomes. This study shows that machine learning approach works well for sentiment analysis of data in particular domain such as movie, product, hotel etc., while lexicon based approach is suitable for short text in micro blogs, tweets, and comments data on web.

A,Nisha Jebaseeliet et al. (2012) Sentiment analysis/opinion mining play vital role to make decision about product /services. Major challenges in opinion mining includes feature weighting which plays a crucial role for good classification.

The above all papers are taken as base work for development of this work. This research work has been developed by using three phases. The first phase of research work is tweets preprocessing, the second phase is sentiment classification and the third phase is performance evaluation of various supervised machine learning algorithms.

Figure 21.2 shows the proposed methodology for this work. The twitter dataset are preprocessed by several steps which includes removal of retweet entities, punctuation, numerals, URL, white space and finally those tweets are converted into lower case. After pre-processing, the second phase is to analyse polarity of tweets using subjectivity lexicon. After classification, the tweets will be converted into Term Document Matrix (TDM). The Classified tweets are sent as input to machine learning algorithms. Finally, the machine learning algorithms performance will be evaluated using accuracy parameter.

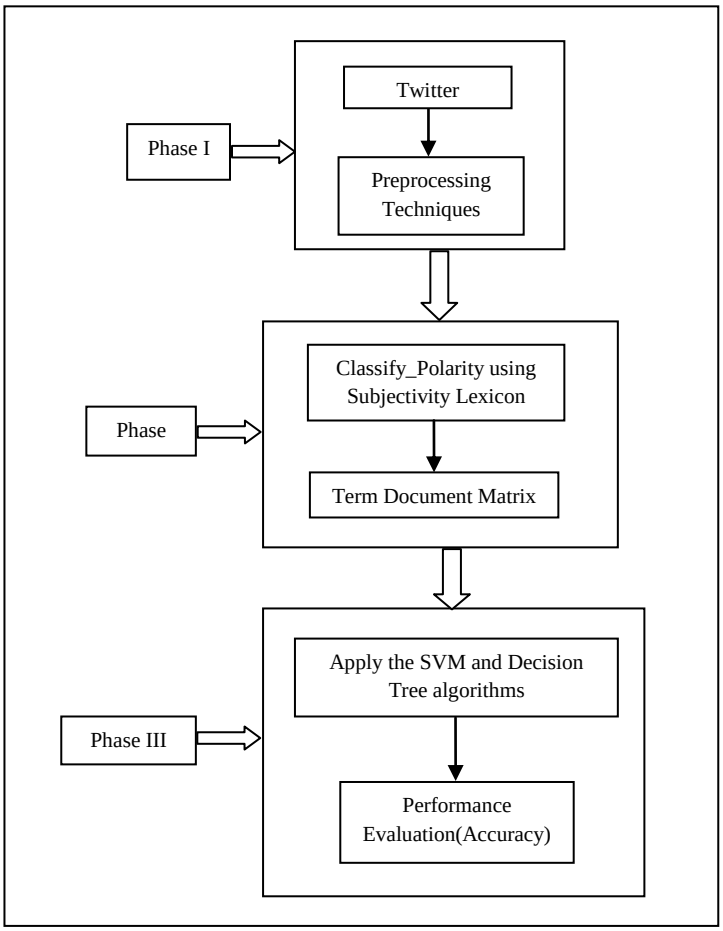

Figure 21.1 Proposed Methodology

The twitter dataset are initially in the first phase needed to be preprocessed by several steps for sentiment classification. Pre-processing steps involved in this proposed methodology are

- **Remove Re-tweet Entities:** Retweet is used in the twitter website to show the tweeting content that has been posted by another user. The format is RT @ username where username is the twitter name of the person who is retweeting.
- **Remove @ people name:** It is necessary to remove the @ <people name> in the tweets which offers better results.
- **Remove Punctuations:** Punctuations in tweets are not necessary for sentiment classification. Punctuations removal yield better results.
- **Remove Numbers:** Numerals are not required for the tweets classification and number removal yields the best performance.
- **Remove URL:** The website links are often attached in the tweet which creates numerous redundancies which are not necessary for sentiment tweets classification.
- **Remove White spaces:** This step is used to remove the Unwanted white space which helps for the tokenization of the tweets.
- **Lowercase:** Converting the tweets to lower case helps the further steps involved in the proposed methodology.

The above Phase – I involves the various pre-processing steps to attain the better performance of sentiment classification. The pre-processed tweets are forwarded as output to the next Phase.

In the second phase, the preprocessed tweets are send as input to the Sentiment package, In the sentiment package two built-in methods are available, which is Naïve Bayesian and Voting method. In that Naïve Bayesian algorithm, two lexicons are available for classify the sentiments, which is subjectivity lexicon and emotion lexicon. The subjectivity lexicon is used to classify the polarity of tweets such as positive, negative and neutral.

Polarity classification is used to classify the tweets into three categories namely positive, negative and neutral. The tweets are classified using Subjectivity lexicon which is developed by Janyce weibe's. The tweets are compared with the subjectivity lexicon and then classified into positive, negative and neutral polarity. After preprocessing, each word is compared with the subjectivity lexicon; if positive words are high in the sentence then the tweet is classified as positive; If negative words are high in the sentence then the tweet is classified as negative; otherwise the tweet is classified as neutral.

After the sentiment classification, the classified tweets are converted into term document matrix, A term-document matrix is a mathematical matrix that describes the frequent words are collected from the documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

The third phase of research work is to apply SVM and Decision tree algorithms to find the best classification algorithm.

Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces(www.wikipedia.com).

Decision Tree

A decision tree is a classification technique which uses the predictive modeling. It always uses "Divide and conquer" strategy where the problem is splitted into search space and subsets .

Definition : *A **decision tree (DT)** is a tree where the root and each internal node is labelled with a question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration*(Dunham, 2006)..

Recursively Partitioning algorithm (RPART) used in decision tree generates both classification and regression models that consists of two stage procedure where the

resulting models represent binary tree. The first stage is building the decision tree by splitting the data into groups and then this process will be applied to each sub-group recursively until it reaches the minimum size or there is no further improvement. The resultant model in this stage is too complex. The second stage contains cross validation to trim back the full tree which is used to estimate the risk of nested sub trees. Pruning (Han.J, 2006) the tree is necessary after building a complete large/complex tree. Given below are the steps involved in pruning the tree; (Therneau).

Performance Evaluation of Algorithms

Evaluation is the process of analyzing the performance of algorithms to find the optimal/best performing algorithm. In this research work, two classification algorithms are used namely Support Vector Machine and Decision Tree. The best performing algorithm for Twitter dataset is interpreted through validation parameter such as precision, recall and accuracy.

Results

This work has been developed using twitter dataset. Twitter is an online social network used to send and read short messages called "tweets". The tweets used to analyze and predict the future directions by public opinion for polling, stock market, products etc. The google self drive car tweets are taken as dataset of this research work. Dataset URL: Twitter Dataset tweets are taken from http://www.crowdflower.com/data-for-everyone ("Twitter Sentiment Analysis: Self-Driving Cars"). Dataset consists of 7156 tweets classified with respect to Google self drive car, Dataset size**:** 1.13 MB.

The proposed research work is undertaken to enhance the Sentiment classification accuracy of twitter data using machine learning algorithms. To attain the objective, various steps have to be performed. The steps are explained in detail with proper results.

Step 1: Preprocessing is one of the necessary step in text mining to remove the unwanted text data.

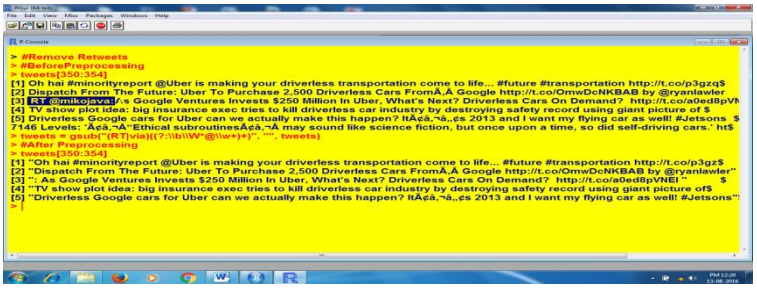

Figure 21.1 Removal of Retweets

Figure 21.2 shows the result of removing the retweets which is RT @username removed from the tweets. Before Preprocessing and after preprocessing tweets are mentioned in the above figure.

Step 2: This step is used to remove the @people name or username of the tweets who posted the tweets in twitter.

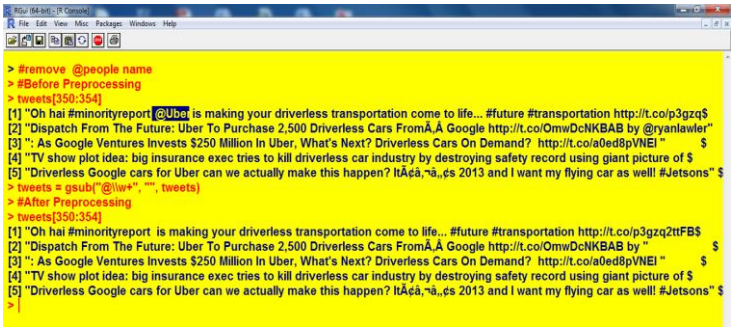

Figure 21.3 Remove the @people name

Figure 21.3 shows the result of @people name removal of tweets, which is @symbol and username of twitter is removed by using gsub function of R tool. Before preprocessing and after preprocessing are mentioned by the above figure.

Step 4: This step is used to remove the punctuation marks in the tweets.

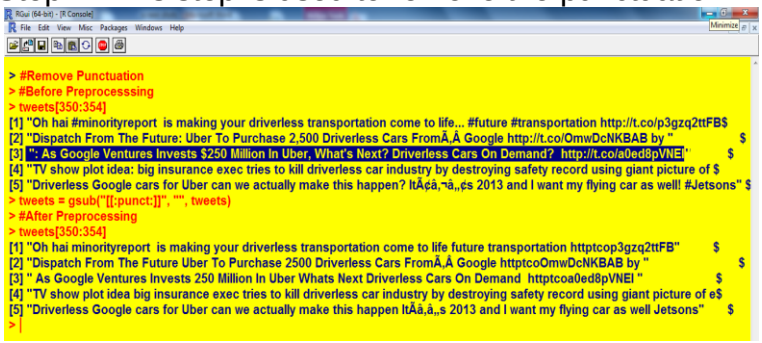

Figure 21.4 Punctuation Removal

Figure 21.4 shows the result of punctuation removal of tweets which is removing the exclamation mark, semicolon, colon, question mark, backslash, dots, etc., Before preprocessing (punctuation removal) and after preprocessed tweets are mentioned in the above figure.

Step 5: This step is used to remove the numbers in the tweets.

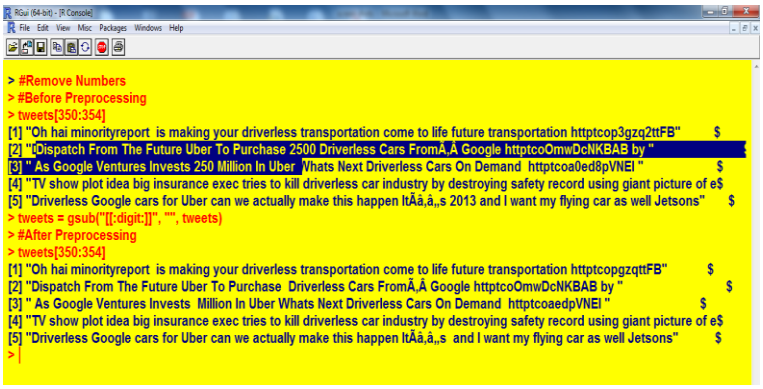

Figure 21.5 Number Removal

Figure 21.5 shows the result of number removal of tweets. Tweets before and after preprocessing mentioned in the above figure.

Step 6: This step is used to remove the URL or html link in the tweets.

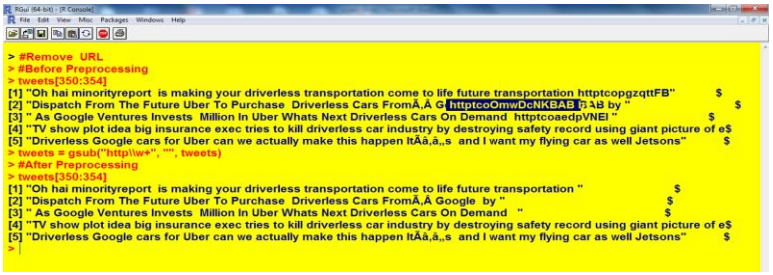

Figure 21.6 URL Removal

Figure 21.6 shows the result of URL or html link removal of tweets, which shows that some users have posted a tweet with some html link. Those links are removed in this step. Tweets before and after preprocessing are mentioned in the above figure.

Step 7: This step is used to convert the tweets from uppercase to lower case.

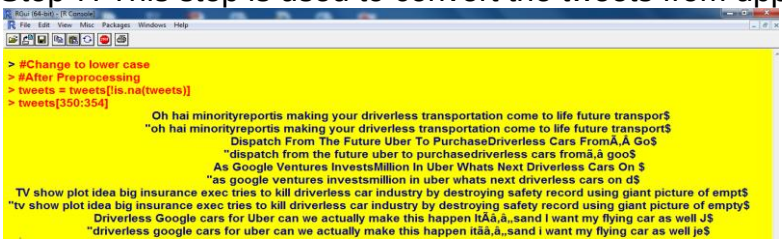

Figure 21.7 Convert from upper case to lower case

Figure 21.7 shows the result of tweets which are converted from upper case to lower case. This step is essential in R tool, which is case sensitive.

Step 8: In this, classifying the polarity using subjectivity lexicon is done; R tool provides sentiment package to classify the sentiments. The emotion lexicon is developed by Janyce weibe's.

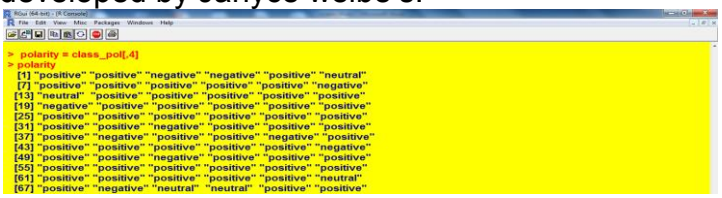

Figure 21.8 Polarity Classification

Figure 8 shows the result of polarity classification, which has classified the tweets into positive, negative and neutral.

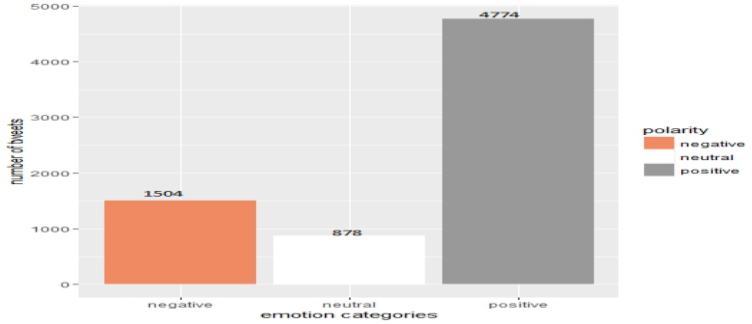

Figure 21.9 Visualization of Polarity Classification

Figure 21.9 shows the result of polarity classification. The Proposed work has taken 7156 tweets dataset for polarity classification. In that 4774 tweets are classified as positive and 1504 as negative tweets, and 878 tweets are classified as neutral. The classification is done using subjectivity lexicon.

```
> recall_accuracy(as.numeric(as.factor(sent_df[51:70, 2])), results2[,"TREE_LABEL"])
[1] 0.65
> recall_accuracy(as.numeric(as.factor(sent_df[51:70, 2])), results1[,"SVM_LABEL"])
[1] 0.9
>
```

Figure 21.10 Performance Analysis of Classification algorithm

Figure 21.10 shows the comparison of two machine learning algorithms, in which Support Vector Machine algorithm achieved 90% accuracy and Decision tree achieved 65% accuracy.

Discussion

The performance metric of two machine learning algorithms are compared. This research work is carried out with subjectivity lexicon and machine learning algorithms. The SVM algorithm achieved better accuracy while compared with decision tree machine learning algorithm. The overall comparison result is shown in figure 11.

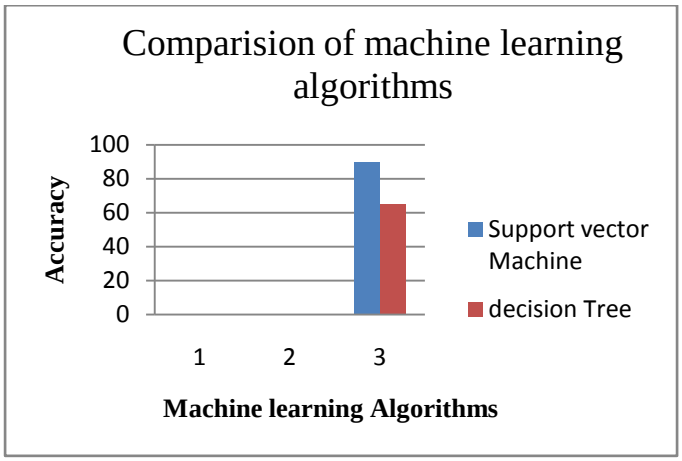

Figure 21.11 Comparison of machine learning algorithms

Conclusion

Sentiment classification is one of the front runners in the text mining techniques which helps business, data scientists and social media. The Social media have become an emerging phenomenon due to the huge and rapid advances in information technology. Social media communications include Face book, twitter, and many others. Twitter is one of the most widely used social media sites. The proposed work has been done on using the twitter dataset, which includes three different phases.

Initially, Twitter Google self drive data are pre-processed by removing re-tweet entities, white space removal, numerals, punctuations, URL or HTML links and user name. After removing unwanted features all tweets are converted into lower case. The pre-processed data are ready to do sentiment classification using subjectivity lexicon. The polarity classified tweets are set into single data frame and it is converted to a Term Document Matrix. The transformed matrix data are divided into training and testing as 70 % and 30 %. There are two algorithms employed to classify and predict the tweets namely Support Vector Machine and Decision Tree and their recall percentage for Twitter Google Self Drive data are 90%, 65%, respectively. From the above inference it is concluded that Support Vector Machine

provides high accuracy when compared with decision tree and also with the existing method.

References

Chetashribhadane, HardiDalal, HeenalDoshi "Sentiment Analysis: Measuring Opinions" Elsevier publications, 2015.

Deebha Mumtaz, Bindiya Ahuja, "A Lexical Approach for Opinion Mining in Twitter", I.J. Education and Management Engineering, 2016, 4, 20-29 Published Online July 2016.

Dunham, Margaret H. Data mining "Introductory and advanced topics", Pearson Education India, 2006.

Farhan Hassan Khan, Saba Bashir, Usman Qamar "TOM Twitter Opinion Mining Frame Work using hybrid classification scheme", Elsevier publication, 2013.

Han, J, Kamber.M, "Data Mining: Concepts and Techniques. Morgan Kauffman", San Francisco ,2006.

NingZhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining" , IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

Therneau, Terry M., and Elizabeth J. Atkinson. An introduction to recursive partitioning using the RPART routines. Technical Report 61. URL http://www. mayo. edu/hsr/techrpt/61. Pdf, 1997.

Twitter Dataset tweets are taken from http://www.crowdflower.com/data-for-everyone ("Twitter Sentiment Analysis: Self-Driving Cars").

Veeramani.S, Karuppusamy.S "A Survey on Sentiment Analysis Technique in Web Opinion Mining" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 3 Issue 8, August 2014.

www.wikipedia.com

www.cis.upenn.edu